HACIA UN NUEVO PARADIGMA EN LA PRODUCCIÓN ESTADÍSTICA: LAS INFRAESTRUCTURAS DE DATOS Y LOS REGISTROS LONGITUDINALES DE POBLACIÓN

Diego Ramiro Fariñas
Director del Instituto de Economía, Geografía y Demografía del Consejo Superior de Investigaciones Científicas

Las sociedades europeas se enfrentan a grandes desafíos debido a los rápidos cambios sociales y económicos. Estos cambios incluyen transformaciones en las formas de familia, en la fecundidad, en la mortalidad y en la longevidad, en las migraciones, y pueden suponer inestabilidad y desigualdad social, así como altas tasas de desempleo. Debido al envejecimiento de la población, todos los países de la UE se enfrentan ahora al desafío de reformar sus sistemas de bienestar. Mientras tanto, el desempleo es alto en muchos países, especialmente en el sur de Europa, lo que se suma a los desafíos causados por el envejecimiento de la población.

Aunque toda esta situación es bien conocida a través de las estadísticas nacionales sobre paro, gasto social, padrón y movimiento natural de la población, así como a través de diversas encuestas, dichos datos no pueden utilizarse para comprender la causa de estas trasformaciones. Para conocer la *causalidad*, necesitamos herramientas que puedan medir esos cambios de forma continua, sin esperar a operaciones censales o grandes encuestas y para ello necesitamos datos longitudinales detallados a nivel individual, con una ventana de observación que abarque un período lo más grande posible de tiempo, que al menos nos permita seguir las biografías individuales dentro del ciclo vital de una generación. Si bien estos datos ya existen en formato digital desde la década de los sesenta en varios países europeos, no existen para muchos otros, y su ventana de observación, cuando existen, se limita a los últimos años del siglo XX y el comienzo del siglo XXI.

Dentro de este escenario de cambios sociales cada vez más rápidos, y de grandes retos demográficos, los institutos de estadística y el mundo de la investigación se enfrentan a un cambio de paradigma en la producción y el uso y explotación de los datos. El desarrollo de los sistemas informáticos ha permitido el uso masivo de información y la creación de infraestructuras de datos que hacen cada vez un uso más eficiente de la información que a la administración ha aportado el ciudadano, permitiendo que en breve la administración pública pueda ofrecer una mayor cantidad y calidad de productos estadísticos. Ya el tamaño de datos que se use no importa sino la capacitación de los investigadores y los estadísticos para su manejo. Estas infraestructuras de datos que serán, o son ya, la columna vertebral de la producción estadística de los países más avanzados en el mundo, debería considerarse como infraestructura básica del Estado, como una más de las infraestructuras que proporcionan los servicios esenciales para la sociedad, y como tal deberían ser cuidadas y dotadas de personal y recursos para ser manejadas y explotadas.

Dentro de este cambio de paradigma, en las últimas décadas, equipos de investigación en toda Europa y Norteamérica, han comenzado a cerrar la brecha entre los datos históricos y contemporáneos, creando nuevas herramientas para comprender y abordar los desafíos sociales causados, por ejemplo, por el envejecimiento de la población y los cambios en las estructuras de empleo. Estos equipos han realizado inversiones a largo plazo en el desarrollo y construcción de registros longitudinales de población y grandes bases de datos de investigación, lo que abre nuevas vías para nuevos enlaces entre diferentes fuentes de datos (como por ejemplo entre datos administrativos y sanitarios). Esos avances metodológicos han dado como resultado la reconstrucción de cientos de miles de cursos de vida individuales y biografías multidimensionales de personas. Estas bases de datos, son el punto de partida para una mejor comprensión de las estabilidades y las transformaciones en nuestras sociedades. El desarrollo de estas bases de datos longitudinales que cubren un período largo de tiempo, junto con el desarrollo de nuevas metodologías de análisis de ciclos de vida y transmisión intergeneracional de características socioeconómicas, demográficas, así como de salud, han llegado a convertir este área de investiga-

ción en una de las más dinámicas en la actualidad en ciencias sociales, humanidades y ciencias médicas, y una de las áreas de investigación en las que podemos esperar un mayor progreso en el futuro.

Por otro lado, hay un movimiento en los institutos y agencias nacionales y subnacionales de estadística, en el que se están desarrollando operaciones estadísticas basadas en la reutilización de datos, bien para mejorar la estadística pública actual, bien para substituir otra serie de operaciones estadísticas con nuevas fuentes de datos, o bien, en el mejor de los casos, y que con más optimismo se sigue en el mundo de la investigación y más frutos puede deparar a la estadística pública, para crear registros longitudinales de población y salud que vinculen datos administrativos recopilados de manera rutinaria. Estos nuevos registros de población se están desarrollando no solo en los países nórdicos, sino también en otros lugares, como Alemania, Países Bajos, Italia, Canadá, Estados Unidos y España. Ejemplos concretos son los registros de población suecos y holandeses, disponibles a nivel nacional, y enlazados en algunos casos a bases de datos que se remontan al siglo XVII. Sin embargo, otros países han desarrollado operaciones ambiciosas para integrar diferentes fuentes de datos administrativos. Por ejemplo, Suiza, donde un consorcio que asocia a la Oficina Federal de Estadística, los Institutos de Medicina Social y Preventiva y el CIGEV de Ginebra, vinculó aproximadamente el 94 por ciento de los 800.000 certificados de defunción recogidos desde 1990 hasta 2008 a los censos de población de 1990 y 2000. El caso escocés, con su Estudio Longitudinal del 5,3% de la población escocesa, que incluye datos del censo de 1991–2011, datos de registro civil, registros de educación y datos de salud y que en breve será enlazado con todo su registro civil desde 1858 a la actualidad a través del proyecto *Digitising Scotland*. El Instituto de Estadística y Cartografía de Andalucía, en colaboración con el CSIC, y gracias al Instituto Nacional de Estadística, con un estudio similar de vinculación de registros en esta región española con la *Base de Datos Longitudinal de Población de Andalucía (BDLPA)*. O el caso de los Estados Unidos, con el *Census Longitudinal Infrastructure Project (CLIP)* que pretende enlazar todos los censos de 1940 a la actualidad creando una infraestructura única de registros longitudinales.

Todas estas bases de datos constan de "big data" multinivel y de múltiples fuentes que incluyen información demográfica, sociológica, intra e intergeneracional a niveles micro, meso y macro, y que permiten el enlace de tres generaciones de trayectorias de vida en, por ejemplo, nacimiento, migración, matrimonio y muerte. Estos registros longitudinales de población representan fuentes de datos increíblemente ricas; entre otras posibilidades, se pueden utilizar para identificar dinámicas domésticas complejas, movilidad social, el estudio del efecto de las condiciones de vida en la infancia en la vida adulta, en su salud y en las desigualdades en la longevidad de las personas mayores, pero también el impacto de las políticas sociales y las intervenciones médicas.

Dos obstáculos para un uso más amplio y extendido de esas grandes bases de datos y registros de población longitudinales son la compleja gestión de datos, la vinculación de conjuntos de datos y la complejidad en la obtención de consistencia en la información de las biografías individuales y las técnicas estadísticas necesarias para su análisis. La unidad de análisis es el curso de la vida individual, la biografía individual, y la investigación longitudinal implica el análisis del flujo de cohortes sucesivas a través de eventos o transiciones definidas y estados o características[1]. A partir de los trabajos pioneros de Cox (1974), el análisis longitudinal ha evolucionado rápidamente y muchas de esas limitaciones se han resuelto[2]. Por otro lado, recientemente se reconoció que muchos datos del curso de la vida son espacialmente ciegos, con información limitada sobre los lugares donde viven las personas. La explosión del análisis de la ciencia de los datos con la reutilización de datos diferentes, difusos y no estructurados, junto con la implementación y el mayor uso de los Sistemas de Información Geográfica, con la creación de infraestructuras de datos espaciales, ha abierto nuevas vías para el uso y análisis de registros poblacionales longitudinales.

En 1946, el Dr. Halbert L. Dunn[3], entonces jefe de la U.S. National Office of Vital Statistics, acuñó el término *"enlace de registros"* y escribió que *"cada persona en el mundo crea un libro de la vida"*. El libro comienza con el nacimiento y termina con la muerte. Sus páginas se componen de los principales acontecimientos de la vida. La vinculación de registros es el nombre que se le da al proceso de ensamblar las páginas del libro en un volumen, y es la base de los registros longitudinales de población. Esperemos que las nuevas Infraestructuras de datos estadísticos se conviertan en un futuro no muy lejano, en las nuevas bibliotecas vivas que contengan los *libros de la vida* de todos nosotros y de nuestros antepasados.

---

[1] Blossfeld, H. P., Hamerle, A., and Mayer, K.U. (1989). *Event History Analysis: Statistical Theory and Application to Social Sciences*, Hillsdale: LEA.

[2] Cox, D. R. (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society*, Series B 34 (2): 187–220. See also Cox, D. R., Oakes, D. (1984). Analysis of Survival Data, New York: Chapman & Hall.

[3] Dunn HL (1946). "Record linkage". *American Journal of Public Health* 36:1412-6.